

# DynaMind: Reconstructing Dynamic Visual Scenes from EEG by Aligning Temporal Dynamics and Multimodal Semantics to Guided Diffusion

Junxiang Liu\*, Junming Lin\*, Jiangtong Li, Jie Li

Tongji University

{junxiang\_liu,jiangtongli}@tongji.edu.cn

## Abstract

Reconstruction dynamic visual scenes from electroencephalography (EEG) signals remains a primary challenge in brain decoding, limited by the low spatial resolution of EEG, a temporal mismatch between neural recordings and video dynamics, and the insufficient use of semantic information within brain activity. Therefore, existing methods often inadequately resolve both the dynamic coherence and the complex semantic context of the perceived visual stimuli. To overcome these limitations, we introduce **DynaMind**, a novel framework that reconstructs video by jointly modeling neural dynamics and semantic features via three core modules: a Regional-aware Semantic Mapper (**RSM**), a Temporal-aware Dynamic Aligner (**TD**A), and a Dual-Guidance Video Reconstructor (**DGVR**). The **RSM** first utilizes a regional-aware encoder to extract multimodal semantic features from EEG signals across distinct brain regions, aggregating them into a unified diffusion prior. In the mean time, the **TD**A generates a dynamic latent sequence, or blueprint, to enforce temporal consistency between the feature representations and the original neural recordings. Together, guided by the semantic diffusion prior, the **DGVR** translates the temporal-aware blueprint into a high-fidelity video reconstruction. On the SEED-DV dataset, **DynaMind** sets a new state-of-the-art (SOTA), boosting reconstructed video accuracies (video- and frame-based) by 12.5 and 10.3 percentage points, respectively. It also achieves a leap in pixel-level quality, showing exceptional visual fidelity and temporal coherence with a 9.4% SSIM improvement and a 19.7% FVMD reduction. This marks a critical advancement, bridging the gap between neural dynamics and high-fidelity visual semantics.

## 1 Introduction

EEG signals, which contain rich human cognitive information (Dahal et al. 2011), have been successfully applied to tasks such as motor imagery (Al-Saegh, Dawwd, and Abdul-Jabbar 2021) and emotion recognition (Liu et al. 2020). As visual stimuli constitute over 60% of sensory input (Scholler et al. 2012), extracting visual information from EEG is crucial for understanding fundamental cognitive processes (Song et al. 2023). Pioneering research, such as Singh et al. (2023) and Wei et al. (2024), has advanced this area, establishing visual reconstruction from EEG as a significant emerging field. Contemporary methods leverage powerful

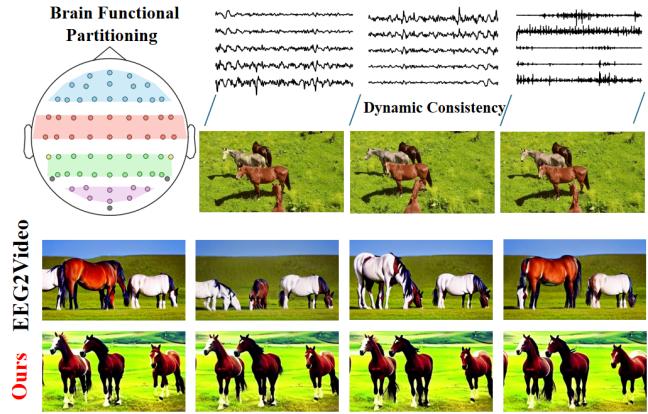


Figure 1: Our approach captures spatio features from diverse brain regions and enforce temporal dynamic consistency between EEG signals and video. This results in generated videos with superior fidelity, temporal coherence, and semantic accuracy compared to EEG2Video.

deep learning models such as CLIP (Radford et al. 2021) and Stable Diffusion (SD) (Rombach et al. 2022). The prevailing framework projects EEG features into the latent space of CLIP to derive semantic representations of the visual stimuli, which then condition generative models for image synthesis. The success of this approach in achieving high image quality and semantic accuracy has established a reliable and efficient generative paradigm. *This success naturally paves the way for the next frontier in EEG based reconstruction, i.e., capturing the dynamic visual world from brain activity to generate coherent video sequences.*

The domain of video reconstruction from EEG, while nascent, presents significant scientific and technical challenges. As illustrated in Fig. 1, distinct functional brain regions exhibit differential patterns of activity when processing complex visual information (Li et al. 2024). Furthermore, neural responses to dynamic visual cues, such as motion velocity and direction, are encoded by distinct cortical areas (Goodale and Milner 1992). However, current research predominantly extracts semantic information from the occipital lobe, often neglecting valuable cognitive information from other brain areas (Zhang et al. 2025). Therefore, by at-

\*These authors contributed equally.

tempting to extract only visual information from EEG (Liu et al. 2024a), these approaches overlook the value of critical multimodal information within the EEG signals themselves. Moreover, existing studies lack dynamic consistency between stimulus representation and the original neural recordings. As shown in Fig. 1, models like EEG2Video (Liu et al. 2024c) often produce reconstructions with intra-frame semantic errors and inter-frame temporal inconsistencies. *Collectively, these limitations present a significant barrier for high-fidelity EEG-to-Video reconstruction.*

To address these challenges, we propose **DynaMind**, a novel framework designed to enrich semantic features using a brain-functional partitioning approach while enforcing dynamic coherence with the video content. The framework accomplishes this task through the synergistic operation of three key components: a Regional-aware Semantic Mapper (**RSM**), a Temporal-aware Dynamic Aligner (**TDA**), and a Dual-Guidance Video Reconstructor (**DGVR**). **RSM** extracts features from distinct functional brain regions, employs multimodal constraints to enrich their semantic diversity, and then aggregates these features into a unified diffusion prior. Meanwhile, **TDA** addresses temporal dynamics by exploiting EEG features to generate a blueprint sequence that enforces dynamic consistency between the neural recordings and video stimulus. Building upon this foundation, **DGVR** leverages the multimodal diffusion prior as semantic guidance to translate the dynamically coherent temporal blueprint into high-fidelity videos exhibiting both temporal consistency and semantic accuracy.

Experiments on the SEED-DV dataset (Liu et al. 2024c) demonstrate that our **DynaMind** framework not only improves direct EEG classification accuracy and the classification accuracy based on its reconstructed videos, but also achieves a leap in pixel-level quality. This is evidenced by a 9.4% improvement in frame-level Structural Similarity (SSIM) and a 19.7% reduction in Fréchet Video Motion Distance (FVMD). Furthermore, we conduct detailed ablation studies to analyze the contributions of different brain regions and the supportive role of multimodal feature integration. Overall, our results mark a significant advancement in the decoding of dynamic visual perception from EEG signals. Our main contributions are summarized as follows:

- We identify and analyze key limitations in EEG-based video reconstruction, namely the underutilization of regional brain information and the poor temporal consistency of existing methods.
- We propose DynaMind, a novel three-module framework designed to improve the dynamic coherence and semantic richness of video reconstructions from EEG signals.
- We validate our framework on the public SEED-DV dataset, establishing a new SOTA in both classification accuracy and reconstruction quality, thus providing a new benchmark for the field.

## 2 Related works

### 2.1 Decoding Visual Information from Brain

A substantial body of research has focused on decoding visual information from physiological signals, uti-

lizing modalities such as functional magnetic resonance imaging (fMRI) (Du et al. 2018), magnetoencephalography (MEG) (Adjamian et al. 2008), and electroencephalography (EEG) (Sun et al. 2023). To accomplish this, a variety of generative models have been explored, prominently featuring variational autoencoders (VAEs) (Kavasidis et al. 2017) and generative adversarial networks (GANs) (Song et al. 2023). More recently, the advent of advanced diffusion models like Stable Diffusion (SD) has spurred significant progress in visual reconstruction (Rombach et al. 2022). Generative schemes employing these models primarily rely on rich embeddings to guide the diffusion process toward specific semantic objectives. Therefore, numerous studies have attempted to derive such semantic information from EEG (Song et al. 2023; Liu et al. 2024a). Its inherent high temporal resolution presents a unique and critical advantage for the more complex task of dynamic video reconstruction.

However, the high signal-to-noise and low spatial resolution ratio of EEG still renders feature extraction difficult. Current methods often attempt to solve this by focusing exclusively on the relationship between EEG and image semantics, overlooking valuable multimodal information inherent in the brain signals themselves. This limitation becomes particularly salient in EEG-to-video generation, where it may lead to videos that are conceptually approximate but exhibit deviations. While pioneering work such as EEG2Video (Liu et al. 2024c) has ventured into this area, its resulting limitations in semantic accuracy and temporal coherence highlight the critical need for further research in the field of video reconstruction from brain signals.

## 2.2 Video Generation Models

Breakthroughs in Text-to-Image (T2I) generation have been largely driven by large-scale multimodal datasets composed of billions of text-image pairs (Nichol et al. 2021; Saharia et al. 2022). To replicate this success in Text-to-Video (T2V) generation, recent studies (Singer et al. 2022; Villegas et al. 2022) have extended these space-only T2I models to the spatio-temporal domain. By training on large-scale text-video datasets such as WebVid-10M (Bain et al. 2021), these models have achieved promising results, laying a foundation for reconstructing visuals from other modalities.

However, prior research in brain signal-based reconstruction has typically relied on holistic semantics from the entire signal to directly guide the generative process. Fundamentally, successful video generation hinges on preserving the continuous motion of consistent objects across frames (Wu et al. 2023). As shown in the 2nd row of Fig. 1, this reliance on holistic guidance leads to reconstructions with significant temporal inconsistencies between consecutive frames.

## 3 Methodology

As shown in Figure 2, our DynaMind framework employs three synergistic modules to translate raw EEG signals into high-fidelity video reconstructions. Regional-aware Semantic Mapper (RSM) generates a semantic diffusion prior by interpreting distinct patterns of regional brain activity. Temporal-aware Dynamic Aligner (TDA) produces a co-

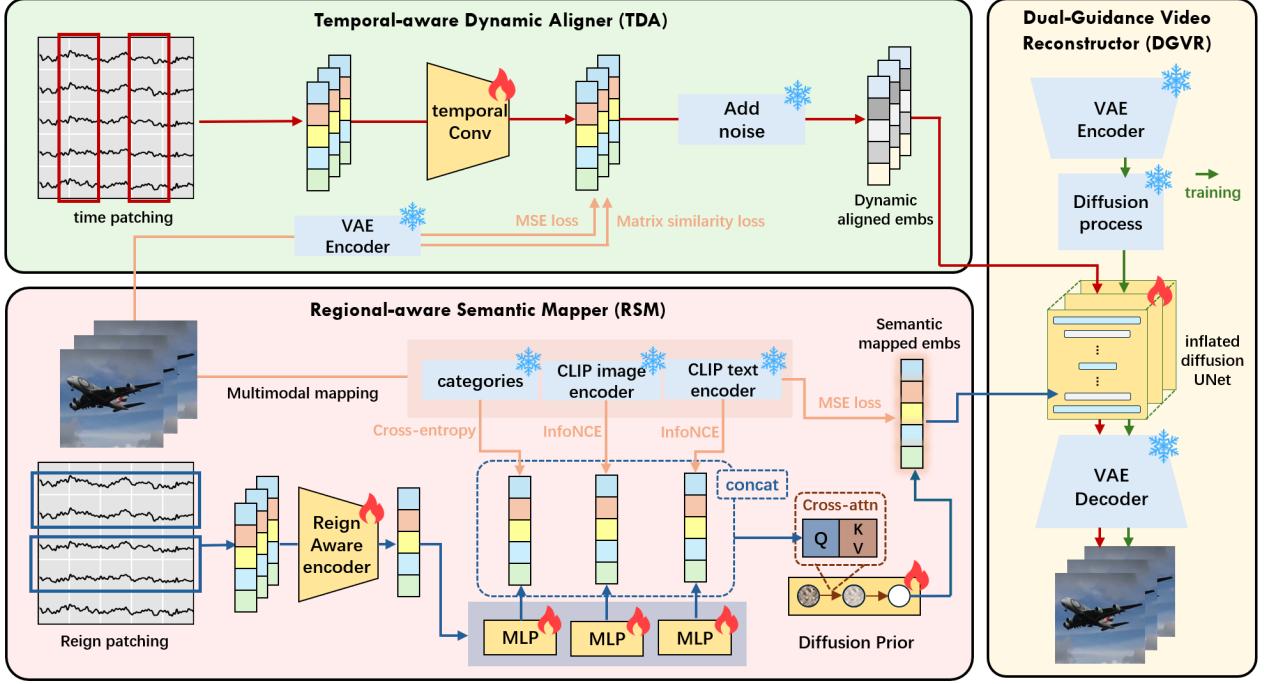


Figure 2: An overview of our DynaMind framework: A dual-guidance architecture for reconstructing high-fidelity videos from EEG signals. Regional-aware Semantic Mapper (RSM) processes EEG signals by brain regions to extract a semantically rich prior, capturing the multimodal content of visual experience. In parallel, Temporal-aware Dynamic Aligner (TDA) processes the global EEG sequence to generate a dynamic temporal blueprint, governing how visual scene evolves over time. Finally, Dual-Guidance Video Reconstructor (DGVR), a latent diffusion model, utilizes the semantic prior for conditional guidance and the temporal blueprint as initial latent variables to synthesize videos, ensuring both semantic accuracy and temporal coherence.

herent temporal blueprint to capture the corresponding visual dynamics. Together, Dual-Guidance Video Reconstructor (DGVR), from a pre-trained video diffusion model, is optimized to accept dual guidance from both the semantic prior and the temporal blueprint, enabling the generated videos with high semantic accuracy and temporal coherence.

### 3.1 Regional-aware Semantic Mapper

**Region-Aware EEG Representation Learning** Human cognitive processes are distributed across functionally specialized networks in the brain. For instance, the occipital lobe is primarily involved in processing visual stimuli, while the temporal and frontal lobes manage more abstract semantic and emotional contexts (Mesulam 1998). Therefore, **RSM** is designed to capture this diversity, leveraging the functional specialization of brain regions to extract semantically rich and varied information from EEG signals.

The process begins by partitioning the multi-channel EEG signals  $\mathbf{E} \in \mathbb{R}^{C \times T}$ , where  $C$  is the channel count and  $T$  is the time points, into  $K$  distinct groups  $\{\mathbf{E}_1, \mathbf{E}_2, \dots, \mathbf{E}_K\}$  based on anatomical brain regions (*i.e.*, frontal, central, temporal, parietal, and occipital lobes). Each regional signal group  $\mathbf{E}_i$  is then processed by a dedicated region-aware encoder. This encoder, consisting of temporal convolutions and a Transformer layer, captures region-specific neural patterns to produce a feature embedding  $f_i$ . These regional embed-

dings  $\{f_1, f_2, \dots, f_K\}$  are fused together to  $\mathbf{F}$ .

$\mathbf{f} = \text{Concat}(f_1, f_2, \dots, f_K)$ , where  $f_i = \text{Enc}_i(\mathbf{E}_i)$ , (1) where each  $\text{Enc}_i$  has independent parameters, allowing it to specialize in its designated brain region’s signals.

**Semantic Enrichment via Multi-Modal Alignment** To enrich these fused embeddings with diverse semantics, we project them into multiple latent spaces for alignment with corresponding multimodal CLIP embeddings. Specifically, we employ a set of mapping networks, denoted as  $M_i$ , to transform the fused EEG representation  $\mathbf{f}$  into the latent spaces of the target multimodal embeddings.

$$\hat{\mathbf{c}}_I = M_{\phi_I}(\mathbf{f}); \hat{\mathbf{c}}_T = M_{\phi_T}(\mathbf{f}); \hat{\mathbf{c}}_Y = M_{\phi_Y}(\mathbf{f}), \quad (2)$$

where  $\phi_I, \phi_T$ , and  $\phi_Y$  are parameters for each mapping network, which are implemented as multi-layer perceptrons.

The alignment between the predicted embeddings ( $\hat{\mathbf{c}}$ ) and ground-truth embeddings ( $\mathbf{c}$ ) is optimized using the InfoNCE loss (Ord, Li, and Vinyals 2018), defined as

$$\mathcal{L}_{\hat{\mathbf{c}} \rightarrow \mathbf{c}} = -\log \frac{\exp(\text{sim}(\hat{\mathbf{c}}, \mathbf{c})/\tau)}{\sum_{j=1}^B \exp(\text{sim}(\hat{\mathbf{c}}, \mathbf{c}_j)/\tau)}, \quad (3)$$

$$\mathcal{L}_{\mathbf{c} \rightarrow \hat{\mathbf{c}}} = -\log \frac{\exp(\text{sim}(\mathbf{c}, \hat{\mathbf{c}})/\tau)}{\sum_{j=1}^B \exp(\text{sim}(\mathbf{c}, \hat{\mathbf{c}}_j)/\tau)}, \quad (4)$$

$$\mathcal{L}_{\text{info}}(\hat{\mathbf{c}}, \mathbf{c}) = \frac{1}{2}(\mathcal{L}_{\hat{\mathbf{c}} \rightarrow \mathbf{c}} + \mathcal{L}_{\mathbf{c} \rightarrow \hat{\mathbf{c}}}). \quad (5)$$

where  $\text{sim}(\cdot, \cdot)$  is cosine similarity and  $\tau$  is a temperature hyperparameter. This symmetric formulation robustly aligns the two embedding spaces. Additionally, a cross-entropy (CE) classification loss is applied to the predicted embedding  $\hat{\mathbf{c}}_Y$  after a linear projection to class logits  $\hat{\mathbf{y}}$ :

$$\mathcal{L}_{\text{category}} = -\frac{1}{N} \sum_{i=1}^N (\mathbf{y}_i \cdot \log(\text{Softmax}(\hat{\mathbf{y}}_i))). \quad (6)$$

where  $\hat{\mathbf{y}}$  and  $\mathbf{y}$  represent the predicted class probabilities and one-hot ground-truth labels, respectively.

**Generation of Semantic Diffusion Prior** The generation of high-fidelity video content hinges on producing effective latent embeddings to condition the subsequent generative model. We observe that embeddings within the CLIP text space provide more effective conditioning for the reconstruction process than the direct EEG-derived features. Therefore, translating the EEG-derived features into the CLIP text embedding space is a critical step for enhancing the final reconstruction quality. Inspired by DALLE-2 (Ramesh et al. 2022), we employ a diffusion prior to perform this cross-modal translation. Specifically, this diffusion prior model learns to map the EEG embedding  $\hat{\mathbf{c}}_{\text{diff}}$ , to a corresponding target embedding within the CLIP text space. For training this prior model, we utilize the same prior loss,  $\mathcal{L}_{\text{prior}}$ , as defined in DALLE-2 (Ramesh et al. 2022).

$$\hat{\mathbf{c}}_{\text{diff}} = \text{Concat}(\hat{\mathbf{c}}_I, \hat{\mathbf{c}}_T, \hat{\mathbf{c}}_Y), \quad (7)$$

$$\mathcal{L}_{\text{prior}} = \mathbb{E}_{t \sim [1, T]} \left[ \left\| P_\theta(\hat{\mathbf{c}}_{\mathbf{T}}^{(t)}, t, \hat{\mathbf{c}}_{\text{diff}}) - \hat{\mathbf{c}}_{\mathbf{T}} \right\|^2 \right]. \quad (8)$$

The total loss is a summation of the InfoNCE losses from image and text, the classification loss and the diffusion loss:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{info}}(\hat{\mathbf{c}}_I, \mathbf{c}_I) + \mathcal{L}_{\text{info}}(\hat{\mathbf{c}}_T, \mathbf{c}_T) + \mathcal{L}_{\text{category}} + \mathcal{L}_{\text{prior}}. \quad (9)$$

### 3.2 Temporal-aware Dynamic Aligner

Unlike the RSM, which focuses on semantic content, the TDA is specifically designed to address the critical challenge of temporal coherence. This module generates a dynamic temporal blueprint to ensure that the motion and flow of the reconstructed video remain synchronized with the underlying temporal dynamics of the EEG signals. Therefore, TDA is vital for mitigating the inter-frame inconsistencies that often plague prior EEG-to-video synthesis methods. To achieve this, the TDA operates in two main stages: the direct generation of a temporal blueprint and its subsequent contrastive alignment with ground-truth video features.

**Extraction of Temporal Blueprint** The TDA receives the raw EEG signal sequence  $E \in \mathbb{R}^{C \times T}$  as input. This sequence is first segmented along the temporal dimension into  $N$  non-overlapping windows,  $\{\mathbf{E}_1^t, \mathbf{E}_2^t, \dots, \mathbf{E}_N^t\}$ , where each window  $\mathbf{E}_i^t \in \mathbb{R}^{C \times \frac{T}{N}}$ . Each temporal segment  $\mathbf{E}_i^t$  is then processed by temporal convolutional network, denoted  $\text{TCN}_{\text{temp}}$ . This architecture is selected for its proficiency in modeling long-range dependencies and is optimized here to capture rhythmic and transitional patterns within the neural

signals. The concatenated outputs from the TCN across all segments form the **dynamic temporal blueprint**,  $\mathbf{H}_{\text{temp}}$ :

$$\mathbf{H}_{\text{temp}} = [\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_N] \quad (10)$$

$$\mathbf{h}_i = \text{TCN}_{\text{temp}}(\mathbf{E}_i^t). \quad (11)$$

This dynamic temporal blueprint is a sequence of latent vectors, where  $\mathbf{h}_i$  encodes the anticipated motion and state changes for the corresponding  $i$ -th segment of the video.

**Contrastive Alignment with Video Features** To ensure the temporal blueprint  $\mathbf{H}_{\text{temp}}$  captures both the content and temporal dynamics of the corresponding video, we align it with video features using a dual objective that enforces both content-wise and structural consistency. First, we define the input features. The temporal blueprint from EEG signal is a feature sequence  $\mathbf{H}_{\text{temp}} \in \mathbb{R}^{N \times D_{\text{eg}}}$ , where  $N$  is the number of frames. The corresponding video is processed by a VAE encoder and a flattening layer to yield a sequence of frame-level visual features  $\mathbf{V} \in \mathbb{R}^{N \times D_{\text{vid}}}$ . Next, both sequences are projected into a shared latent space of dimension  $D_{\text{latent}}$  using distinct projection heads,  $P_H$  and  $P_V$ :

$$\mathbf{Z}_H = P_H(\mathbf{H}_{\text{temp}}) \in \mathbb{R}^{N \times D_{\text{latent}}}, \quad (12)$$

$$\mathbf{Z}_V = P_V(\mathbf{V}') \in \mathbb{R}^{N \times D_{\text{latent}}}. \quad (13)$$

With the features projected into this common space, we apply two alignment losses. To enforce direct, frame-by-frame content similarity, we compute the Mean Squared Error (MSE) between the two projected sequences:

$$\mathcal{L}_{\text{HV}} = \frac{1}{N} \sum_{i=1}^N \|(\mathbf{Z}_H)_i - (\mathbf{Z}_V)_i\|^2. \quad (14)$$

Moreover, to align the relational dynamics between frames, we introduce a structural loss. This loss compares the intramodality similarity matrices,  $\mathbf{S}_H \in \mathbb{R}^{N \times N}$  and  $\mathbf{S}_V \in \mathbb{R}^{N \times N}$ , which capture the internal structure of each sequence. The matrices are computed using cosine similarity, e.g.,  $(\mathbf{S}_H)_{ij} = \text{cos-sim}((\mathbf{Z}_H)_i, (\mathbf{Z}_H)_j)$ . The loss then minimizes the MSE between these two structural representations:

$$\mathcal{L}_{\text{Struct}} = \frac{1}{N^2} \|\mathbf{S}_H - \mathbf{S}_V\|_F^2. \quad (15)$$

The final loss for the TDA module is the sum of the content and structural alignment losses:

$$\mathcal{L}_{\text{TDA}} = \mathcal{L}_{\text{HV}} + \mathcal{L}_{\text{Struct}}. \quad (16)$$

### 3.3 Dual-Guidance Video Reconstructor (DGVR)

The **DGVR** is responsible for synthesizing the final high-fidelity video. Adopting the Tune-A-Video technique (Wu et al. 2023) as a foundation, our model is conditioned on both the semantic diffusion prior from the **RSM** and the temporal blueprint from the **TDA**. This dual-guidance mechanism is key to generating videos that are both semantically accurate and temporally coherent.

Instead of initiating the reverse diffusion process from pure Gaussian noise, we structure the initial latent variable  $\mathbf{x}_T$  using the temporal blueprint  $\mathbf{H}_{\text{temp}}$ . This ensures that

Method	40-c top-1	40-c top-5	9-c top-1	9-c top-3	Color	Fast/Slow	Numbers	Human Face	Human
Chance level	2.50	12.50	11.11	33.33	20.57	50.00	65.64	62.25	71.43
ShallowNet	5.59/2.27*	16.93/4.66*	21.40/1.96*	49.62/2.34*	27.00/2.09*	56.62/1.77*	66.15/0.89	64.87/1.54	73.21/1.52
DeepNet	4.56/1.52*	14.30/3.25*	20.27/1.25*	48.06/1.59*	26.37/1.95*	55.42/0.59*	65.71/0.24	61.58/3.93	72.86/0.40
EEGNet	4.64/0.86*	14.25/1.87*	19.63/0.81*	47.04/1.45*	25.46/1.31*	51.99/2.00	64.67/0.60	61.37/1.31	72.38/0.98
Conformer	4.93/1.57*	15.36/4.44*	20.92/0.98*	49.25/1.49*	<b>27.53/1.37*</b>	55.02/0.83*	65.73/0.26	64.96/1.14	73.00/0.85
TSCConv	4.92/0.99*	15.05/2.31*	20.00/1.01*	47.76/1.51*	26.89/1.83*	55.32/0.99*	65.39/0.41	64.39/1.47	72.68/0.67
GLMNet	6.20/3.02*	17.75/4.24*	21.93/1.87*	50.01/2.52*	27.33/1.45*	<b>57.35/1.98*</b>	66.21/0.91	65.10/1.45	73.34/1.31
<b>Ours</b>	<b>8.27/2.85*</b>	<b>22.38/4.80*</b>	<b>22.73/3.07*</b>	<b>51.93/4.73*</b>	27.43/1.11*	56.44/1.51*	<b>67.56/0.94</b>	<b>82.47/0.81*</b>	<b>73.64/1.70</b>

Table 1: Average classification accuracy (%) and std across all subjects with different EEG classifiers on different tasks. Chance level is the percentage of the largest class. The star symbol (\*) represents the result is above chance level with statistical significance (two-sample t-test:  $p < 0.05$ ).

# Classes	Method	Video-based			Frame-based		
		Semantic-level		Pixel-level	Semantic-level		Pixel-level
		2-way	40-way	FVMD↓	2-way	40-way	SSIM↑
10	Ours	0.847±0.01	0.394±0.03	1601.84±0.02	0.833±0.02	0.308±0.01	0.309±0.02
	EEG2Video	0.852±0.02	0.340±0.01	1977.13±0.01	0.798±0.03	0.232±0.02	0.300±0.03
20	Ours	0.835±0.01	0.345±0.01	1587.91±0.01	0.818±0.02	0.277±0.02	0.290±0.02
	EEG2Video	0.813±0.02	0.273±0.03	1960.20±0.03	0.785±0.04	0.184±0.02	0.242±0.03
30	Ours	0.833±0.02	0.309±0.01	1661.06±0.03	0.805±0.02	0.254±0.01	0.293±0.01
	EEG2Video	0.794±0.02	0.209±0.05	2016.67±0.04	0.785±0.04	0.180±0.02	0.228±0.04
40	Ours	0.828±0.02	0.284±0.02	1637.55±0.01	0.807±0.03	0.241±0.01	0.280±0.01
	EEG2Video	0.798±0.03	0.159±0.01	2038.27±0.02	0.774±0.02	0.138±0.01	0.256±0.03

Table 2: Reconstruction results across varying class numbers, evaluated on semantic/pixel metric at the video/frame level.

the foundational spatio-temporal structure of the video is aligned with the neural dynamics from the outset. Specifically, we first project the blueprint to a latent representation  $\mathbf{z}_B$  and define the initial latent as:

$$\mathbf{x}_T = \mathcal{E} + \alpha \cdot \mathcal{U}(\mathbf{z}_B), \quad (17)$$

where  $\mathcal{E} \sim \mathcal{N}(0, \mathbf{I})$  is a random Gaussian noise tensor,  $\mathcal{U}(\cdot)$  is an upsampling network matching the blueprint’s dimensions to the latent space, and  $\alpha$  is a hyperparameter controlling the temporal guidance strength. During each step of the reverse diffusion process (from  $t = T$  down to 1), the semantic diffusion prior,  $\hat{\mathbf{c}}_{diff}$ , is injected into the model’s U-Net architecture via cross-attention layers. This conditioning ensures that the objects, scenes, and overall context of the generated frames are semantically consistent with the information decoded from the various brain regions.

The diffusion model’s U-Net, denoted  $\epsilon_\theta$ , is trained to predict the noise  $\epsilon$  from a noisy latent  $\mathbf{x}_t$  at a given timestep  $t$ . The prediction is conditioned on  $\mathbf{x}_t$ ,  $t$ , the semantic diffusion prior,  $\hat{\mathbf{c}}_{diff}$ , and the projected temporal blueprint  $\mathbf{z}_B$ . The objective function is the standard diffusion loss:

$$\mathcal{L}_{DGVR} = \mathbb{E}_{\mathbf{x}_0, \epsilon, t} \left[ \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \hat{\mathbf{c}}_{diff}, \mathbf{z}_B)\|^2 \right]. \quad (18)$$

where  $\mathbf{x}_t = \sqrt{\bar{\alpha}_t} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$ , and  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ .

After the iterative denoising process is complete, a final decoder module transforms the clean latent representation

$\mathbf{x}_0$  into the pixel-space video. By leveraging both dynamic and semantic guidance, our model reconstructs videos with superior fidelity, coherence, and contextual accuracy.

## 4 Experiments

### 4.1 Experimental Setup

**Dataset and Baselines.** Our experiments are conducted on the **SEED-DV** dataset (Liu et al. 2024c), which comprises EEG signals from 20 subjects viewing video clips across 40 visual conceptual categories. To validate the effectiveness of our model, we benchmark its performance against comprehensive baseline methods, including both foundational EEG models (ShallowNet (Schirrmeister et al. 2017) and EEGNet (Lawhern et al. 2018)), and SOTA approaches (Conformer (Song et al. 2022) and GLMNet (Liu et al. 2024c)).

**Implementation and Evaluation.** Our model is composed of three core modules. The RSM aligns regional EEG embeddings with semantic features from a pre-trained CLIP ViT-L/14 model. In parallel, the TDA extracts a temporal blueprint from the global EEG dynamics. Finally, the DGVR fine-tunes a Stable Diffusion v1.4 model to synthesize 6-frame video clips at  $512 \times 288$  resolution (3 fps), conditioned on the semantic and temporal guidance from the other modules. We follow the evaluation protocol established by EEG2Video (Liu et al. 2024c) to ensure fair comparison.

Method	Classification-based		Video-based			Frame-based		
	40-c top-1	9-c top-1	Semantic-level	Pixel-level	Semantic-level	Pixel-level		
			2-way	40-way	FVMD↓	2-way	40-way	SSIM↑
<b>Ours</b>	<b>8.27</b>	<b>22.73</b>	<b>0.828</b>	<b>0.284</b>	1637.55	0.807	<b>0.241</b>	<b>0.280</b>
<i>Brain Regions</i>								
w/o Frontal	7.67	22.02	0.813	0.264	1677.42	0.793	0.220	0.267
w/o Parietal	7.65	22.01	0.812	0.263	<b>1631.21</b>	0.796	0.221	0.267
w/o Occipital	6.73	20.93	0.807	0.239	1712.11	0.786	0.189	0.253
w/o Temporal	7.22	21.57	0.809	0.252	1658.89	0.790	0.205	0.260
<i>Features</i>								
w/o Image	8.08	22.44	0.825	0.281	1649.77	0.807	0.235	0.277
w/o Text	7.56	21.87	0.811	0.252	1703.48	0.793	0.212	0.270
w/o Category	8.11	22.45	0.820	0.276	1644.07	0.802	0.232	0.278
<i>Consistency</i>								
w/o Consistency	-	-	0.825	0.279	1916.91	<b>0.810</b>	0.240	<b>0.280</b>

Table 3: Quantitative results of the ablation study on the key components of our model. The performance drop after removing individual brain regions, features, or the consistency module validates the effectiveness of each component.

Our assessment includes 1) Frame-Level Quality: The Structural Similarity Index (SSIM) (Wang et al. 2004) is used to measure pixel-level fidelity. N-way top-K classification accuracy based on CLIP features is used to assess semantic correctness; 2) Video-Level Dynamics: N-way top-K accuracy based on VideoMAE features is used to evaluate dynamic semantic coherence. The Frame-wise Video Motion Distance (FVMD) (Liu et al. 2024b) is used to quantify temporal consistency. **More details about the implementation and evaluation are in Appendix.**

## 4.2 Main Results

**Classification Accuracy.** Detailed results are presented in Tab. 1. The experimental outcomes demonstrate that our proposed model surpasses existing baselines across the majority of metrics, thereby establishing a new SOTA performance. In the most challenging 40-class task, our model achieves a Top-1 accuracy of 8.27%, significantly outperforming the runner-up, GLMNet (6.20%). This performance advantage is maintained in the 9-class task, where our model again achieves the highest Top-1 accuracy of 22.73%. This consistent performance margin suggests that our model captures richer and more discriminative semantic details from the EEG signals. Our model’s capabilities are particularly evident in fine-grained attribute recognition. For the “**Human Face**” task, it achieves an accuracy of **82.47%**, a substantial margin of over 17 percentage points above the next-best method (GLMNet at 65.10%). This improvement provides robust evidence of our model’s enhanced capacity for decoding specific, high-level semantic concepts. Furthermore, our model secured leading results in other fine-grained tasks, including “Numbers” and “Human”. In summary, these results provide strong quantitative support for our model’s effectiveness, verifying its ability to learn highly discriminative and semantically meaningful features for the subsequent

high-fidelity video reconstruction.

**Video Reconstruction.** We quantitatively compare our model with the EEG2Video baseline across tasks with varying class numbers (10 to 40). Tab. 2 details the evaluation, which assesses video/frame levels across semantic/pixel dimensions. Our method significantly outperforms the EEG2Video baseline in all evaluated settings. Semantically, our model’s advantage grows with task difficulty as the number of classes increases from 10 to 40. For instance, in the challenging 40-class setting, our model’s video semantic accuracy reaches **0.284**, far surpassing the baseline’s 0.159. Similarly, our frame-based semantic accuracy (**0.241**) substantially exceeds the baseline’s performance (0.138). This result demonstrates our model provides more precise semantic guidance during the generation process. At the pixel level, FVMD assesses video quality and temporal coherence, where lower is better. Our model achieves an FVMD of **1637.55** in the 40-class task, markedly lower than the baseline’s 2038.27. This lower FVMD indicates superior motion smoothness and inter-frame consistency, directly validating the efficacy of our TDA module. Furthermore, a higher SSIM indicates our generated frames possess greater structural fidelity. In conclusion, these quantitative results verify the superiority of our dual-guidance framework across both semantic and pixel dimensions. Our model sets a new performance benchmark for EEG-based video reconstruction in both content accuracy and visual fidelity.

## 4.3 Ablation Study

We conduct ablation studies to validate each key component of our model in Tab. 3. Starting from our full model, we evaluate the contribution of each component by removing it and observing the impact on performance.

**Effect of Brain Regions.** Among these regions, the Occipital Lobe has the most significant impact on performance,

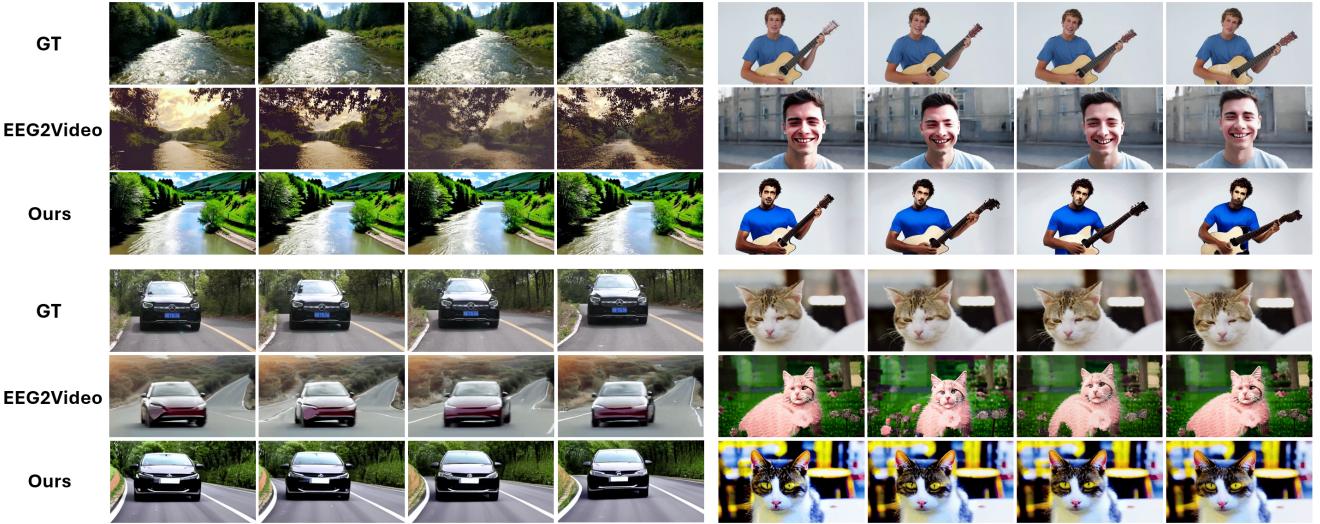


Figure 3: Video reconstruction results of ours and EEG2Video compared against the ground truth for diverse visual concepts. The results of EEG2Video are from their openly released samples.

which aligns with its primary role in visual processing. Nevertheless, removing signals from any single brain region, including the frontal, parietal, or temporal lobes, also degrades performance. This result underscores the collective and distributed contribution of all regional features.

**Effect of Features.** Ablating different modal features reveals that removing the **text feature (w/o Text)** causes the most significant drop in semantic and classification performance, underscoring its critical role. The performance degradation from removing image or category features also validates the effectiveness of our multimodal fusion strategy.

**Effect of Consistency.** Removing the structural consistency loss (**w/o Consistency**) significantly worsens the FVMD score (from 1637.55 to 1916.91). This sharp increase confirms that this objective plays a key role in enhancing the temporal coherence of the generated video.

#### 4.4 Visualization

The qualitative comparison results clearly demonstrate our model’s significant advantage in semantic fidelity. A particularly illustrative case is the reconstruction of the “cat” in the bottom right. Our model successfully generates a cat with realistic morphology and texture, exhibiting high fidelity to the ground truth. In stark contrast, the EEG2Video baseline fails to capture the correct semantics entirely, generating a pink, morphologically distorted object that suffers from severe content hallucination. This superiority is also evident in other examples: the “car” reconstructed by our model features a well-defined contour and accurate coloration, while the “river” scene is rendered with vivid colors and rich details. Conversely, the baseline’s outputs are generally plagued by blurriness, distortion, and color deviation.

Beyond the accuracy of static content, our model also exhibits superior temporal coherence. In the dynamic scenes of the “river” and “car,” the videos we generate feature smooth transitions and stable morphology between frames, signif-

icantly reducing the flickering and warping artifacts common in the baseline model. This result empirically validates the efficacy of our **Temporal Dynamic Aligner (TDA)** in maintaining dynamic consistency. Furthermore, when handling complex scenes containing a “human” and specific “objects” like a guitar, our model accurately reconstructs the core scene composition and human poses, whereas the baseline struggles to produce coherent and meaningful imagery.

In summary, these visual results intuitively corroborate the superior performance of our model. By effectively combining rich semantic guidance with precise temporal dynamic constraints, our model is capable of generating high-quality videos that are semantically accurate, visually detailed, and temporally coherent, marking a significant improvement over existing methods.

## 5 Conclusion

This work addresses key challenges in EEG-to-video reconstruction: suboptimal use of distributed neural information, and a lack of dynamic temporal coherence. Therefore, we propose DynaMind, a new framework that reconstructs dynamic visual scenes by integrating multimodal semantics with temporal dynamics. DynaMind leverages three core modules: a Region-aware Semantic Mapper for extracting region-aware multimodal semantic priors, a Temporally-aware Dynamic Aligner for generating dynamic coherent temporal blueprint, and a Dual-guided Video Reconstructor for synthesizing the final video with both semantic and dynamic guidance. Experiments on the SEED-DV dataset show that DynaMind achieves state-of-the-art results, outperforming existing methods in both direct EEG classification and reconstructed video quality with notable improvements in semantic accuracy, pixel fidelity, and temporal consistency. Ablation studies confirm the necessity of each component, signifying a major advance in decoding visual experiences from non-invasive brain recordings.

## References

- Adjamian, P.; Hadjipapas, A.; Barnes, G. R.; Hillebrand, A.; and Holliday, I. E. 2008. Induced Gamma activity in primary visual cortex is related to luminance and not color contrast: An MEG study. *Journal of Vision*, 8(7): 4–4.
- Al-Saegh, A.; Dawwd, S. A.; and Abdul-Jabbar, J. M. 2021. Deep learning for motor imagery EEG-based classification: A review. *Biomedical Signal Processing and Control*, 63: 102172.
- Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1728–1738.
- Dahal, N.; Nandagopal, N.; Nafalski, A.; and Nedic, Z. 2011. Modeling of cognition using EEG: a review and a new approach. In *TENCON 2011-2011 IEEE Region 10 Conference*, 1045–1049. IEEE.
- Du, C.; Du, C.; Huang, L.; and He, H. 2018. Reconstructing perceived images from human brain activities with Bayesian deep multiview learning. *IEEE transactions on neural networks and learning systems*, 30(8): 2310–2323.
- Goodale, M. A.; and Milner, A. D. 1992. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1): 20–25.
- Kavasidis, I.; Palazzo, S.; Spampinato, C.; Giordano, D.; and Shah, M. 2017. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, 1809–1817.
- Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.
- Lawhern, V. J.; Solon, A. J.; Waytowich, N. R.; Gordon, S. M.; Hung, C. P.; and Lance, B. J. 2018. EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of neural engineering*, 15(5): 056013.
- Li, D.; Wei, C.; Li, S.; Zou, J.; Qin, H.; and Liu, Q. 2024. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*.
- Liu, H.; Hajialigol, D.; Antony, B.; Han, A.; and Wang, X. 2024a. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pre-training and multi-view transformer. *arXiv preprint arXiv:2405.02165*.
- Liu, J.; Qu, Y.; Yan, Q.; Zeng, X.; Wang, L.; and Liao, R. 2024b. Fr\`echet Video Motion Distance: A Metric for Evaluating Motion Consistency in Videos. *arXiv preprint arXiv:2407.16124*.
- Liu, S.; Wang, X.; Zhao, L.; Zhao, J.; Xin, Q.; and Wang, S.-H. 2020. Subject-independent emotion recognition of EEG signals based on dynamic empirical convolutional neural network. *IEEE/ACM transactions on computational biology and bioinformatics*, 18(5): 1710–1721.
- Liu, X.-H.; Liu, Y.-K.; Wang, Y.; Ren, K.; Shi, H.; Wang, Z.; Li, D.; Lu, B.-L.; and Zheng, W.-L. 2024c. EEG2video: Towards decoding dynamic visual perception from EEG signals. *Advances in Neural Information Processing Systems*, 37: 72245–72273.
- Mesulam, M.-M. 1998. From sensation to cognition. *Brain: a journal of neurology*, 121(6): 1013–1052.
- Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.
- Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.
- Schirrmeyer, R. T.; Springenberg, J. T.; Fiederer, L. D. J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; and Ball, T. 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, 38(11): 5391–5420.
- Scholler, S.; Bosse, S.; Treder, M. S.; Blankertz, B.; Curio, G.; Muller, K.-R.; and Wiegand, T. 2012. Toward a direct measure of video quality perception using EEG. *IEEE transactions on Image Processing*, 21(5): 2619–2629.
- Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.
- Singh, P.; Pandey, P.; Miyapuram, K.; and Raman, S. 2023. EEG2IMAGE: image reconstruction from EEG brain signals. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Song, Y.; Liu, B.; Li, X.; Shi, N.; Wang, Y.; and Gao, X. 2023. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*.
- Song, Y.; Zheng, Q.; Liu, B.; and Gao, X. 2022. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31: 710–719.

Sun, J.; Li, M.; Chen, Z.; Zhang, Y.; Wang, S.; and Moens, M.-F. 2023. Contrast, attend and diffuse to decode high-resolution images from brain activities. *Advances in Neural Information Processing Systems*, 36: 12332–12348.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Video-mae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093.

Villegas, R.; Babaeizadeh, M.; Kindermans, P.-J.; Moraldo, H.; Zhang, H.; Saffar, M. T.; Castro, S.; Kunze, J.; and Erhan, D. 2022. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wei, Y.; Cao, L.; Li, H.; and Dong, Y. 2024. Mb2c: Multimodal bidirectional cycle consistency for learning robust visual neural representations. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8992–9000.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7623–7633.

Zhang, K.; He, L.; Jiang, X.; Lu, W.; Wang, D.; and Gao, X. 2025. CognitionCapturer: Decoding Visual Stimuli From Human EEG Signal With Multimodal Information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 14486–14493.

## Appendix

This appendix provides additional details, analyses, and experiments to support our main findings. The contents are organized as follows: Appendix A provides detailed descriptions of our experimental settings; Appendix B provides detailed descriptions of our evaluation metrics; Appendix C provides detailed descriptions about our compared baselines; Appendix D provides extra visualization about successful and failure cases.

## A Experimental Setup

### A.1 Dataset

We use the public SEED-DV dataset (Liu et al. 2024c), which comprises 40 distinct visual concepts. These concepts fall into nine higher-level categories, including Land Animal, Water Animal, Plant, Exercise, Human, Natural Scene, Food, Musical Instrument, and Transportation. For each concept, the dataset provides 35 video clips along with the corresponding EEG signals recorded from 20 subjects.

### A.2 Implementation Details

Our model comprises three core modules, each with a distinct implementation and training protocol:

- **Regional-aware Semantic Mapper (RSM):** EEG data is parcellated into  $K = 4$  neuroscientifically-informed brain regions. Each region’s data is processed by independent layers to produce 512-dimensional embeddings, which are fused and projected into a unified 1024-dimensional feature. This feature is aligned with the text and image feature space of a pre-trained CLIP ViT-L/14 model. The RSM encoder and alignment mappings are jointly trained for 300 epochs (learning rate  $1 \times 10^{-5}$ ). A separate DALLE-inspired prior module is then trained for 1000 epochs (learning rate  $2 \times 10^{-5}$ ).
- **Temporal-aware Dynamic Aligner (TDA):** EEG signals are segmented into 6-step windows. An encoder processes these windows to generate a temporal feature sequence. The module is trained for 300 epochs (learning rate  $1 \times 10^{-5}$ ) to align its output with VAE latent features derived from the ground-truth video via DDIM inversion.
- **Dual-Guidance Video Reconstructor (DGVR):** We fine-tune a Stable Diffusion v1.4 model for 200 epochs with a learning rate of  $3 \times 10^{-5}$ , following the protocol in (Wu et al. 2023). Training data consists of 6-frame video clips at a  $512 \times 288$  resolution and 3 fps. All models are trained on an NVIDIA A800 GPU using a cosine annealing learning rate scheduler.

## B Evaluation Metrics

Our evaluation follows the protocol established by EEG2Video (Liu et al. 2024c) to ensure a fair comparison. First, we assess frame-level quality via two metrics: the Structural Similarity Index (SSIM) (Wang et al. 2004) for pixel fidelity, and N-way top-K accuracy using a CLIP-based classifier (Radford et al. 2021) for semantic content. Acknowledging that static frame analysis cannot capture temporal dynamics, we introduce complementary video-level

metrics. We employ a VideoMAE (Tong et al. 2022) classifier, pre-trained on the Kinetics-400 dataset (Kay et al. 2017), to evaluate the dynamic semantic coherence of the generated video clips. Furthermore, we use the Frame-wise Video Motion Distance (FVMD) (Liu et al. 2024b) metric to assess the temporal consistency and motion smoothness of the reconstructed videos.

**Structural Similarity Index (SSIM)** SSIM assesses the pixel-level fidelity of a reconstructed frame by comparing its structural information with a ground-truth frame. It is a perception-based model that considers changes in luminance, contrast, and structure. For two image windows  $x$  (ground truth) and  $y$  (reconstruction), SSIM is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (19)$$

where  $\mu_x$  and  $\mu_y$  are the average pixel values,  $\sigma_x^2$  and  $\sigma_y^2$  are the variances, and  $\sigma_{xy}$  is the covariance of  $x$  and  $y$ .  $C_1 = (k_1 L)^2$  and  $C_2 = (k_2 L)^2$  are stabilization constants, where  $L$  is the dynamic range of pixel values. The final score is the mean SSIM over all windows.

**Frame-wise Video Motion Distance (FVMD)** FVMD evaluates the temporal consistency and motion realism among frames. It is analogous to Fréchet Inception Distance but is applied to motion features extracted from consecutive frames. First, motion features (e.g., from optical flow) are extracted for all pairs of adjacent frames from both ground-truth videos and generated videos. These sets of features are then modeled as multivariate Gaussian distributions. FVMD is the Fréchet distance between these two distributions:

$$\text{FVMD} = \|\mu_r - \mu_g\|_2^2 + \text{Tr} \left( \Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2} \right). \quad (20)$$

where  $(\mu_r, \Sigma_r)$  and  $(\mu_g, \Sigma_g)$  are the mean and covariance of the motion features from the real and generated videos, respectively, and  $\text{Tr}(\cdot)$  is the trace of a matrix. A lower FVMD score indicates more realistic and coherent motion.

**N-way top-K Accuracy on Frames and Videos** This metric evaluates the semantic accuracy of a single reconstructed frame or whole video. For a given frame or video, its class is predicted from a set of  $N$  randomly chosen classes (the ground-truth class plus  $N - 1$  “distractor” classes). The prediction is considered correct if the ground-truth class is within the top- $K$  predicted probabilities. For a test set of size  $M$ , the accuracy is:

$$\text{Acc}_{N,K} = \frac{1}{M} \sum_{i=1}^M \mathbb{I}(y_i \in \text{TopK}(\hat{P}_i(S_N))). \quad (21)$$

where  $y_i$  is the ground-truth label for sample  $i$ ,  $S_N$  is the set of  $N$  classes,  $\hat{P}_i(S_N)$  are the predicted probabilities for classes in  $S_N$ ,  $\text{TopK}(\cdot)$  returns the set of  $K$  classes with the highest probabilities, and  $\mathbb{I}(\cdot)$  is the indicator function. For both frame and video-level evaluations, we employ 40-way and 2-way top-1 accuracy to assess the semantic quality of the reconstructed content. For the frame-level classifier, we follow the protocol in Liu et al. (2024c) and fine-tune

a pre-trained CLIP encoder on the ground-truth frames for the 40-way visual concept classification task. Similarly, for the video-level classifier, a pre-trained VideoMAE model is fine-tuned on the ground-truth video clips, consistent with the same evaluation protocol (Liu et al. 2024c).

## C Baselines

We evaluate our model’s performance against a comprehensive set of strong and representative baselines from the field of EEG analysis. These selected baselines range from foundational CNN architectures to recent state-of-the-art models.

**ShallowNet and DeepNet** are two foundational CNN architectures for EEG decoding proposed by Schirrmeister et al. (2017). **ShallowNet** employs both temporal and spatial convolutional layers, an architecture effective for decoding frequency power information. In contrast, **DeepNet** is a deeper, VGG-inspired network designed to learn complex features through stacked convolutional blocks.

**EEGNet**, from Lawhern et al. (2018), is a compact and efficient CNN architecture designed for EEG-based brain-computer interfaces (BCIs). The model uses depthwise and separable convolutions to extract temporal and spatial features with fewer parameters, which enhances its generalizability across different BCI paradigms.

**Conformer**, from Song et al. (2022), is an architecture for EEG signal processing that combines the strengths of both CNNs and Transformers. The model employs convolutional layers to extract local features and a multi-head self-attention mechanism to model global correlations, an approach proven effective for various EEG decoding tasks.

**TSConv.** The Temporal-Spatial Convolutional network (TSConv), from Song et al. (2023), is an advanced CNN-based model for EEG analysis. It uses dedicated convolutional blocks to explicitly learn features along both temporal and spatial dimensions, which enhances its ability to extract discriminative information.

**GLMNet**, from EEG2Video Liu et al. (2024c), is a recent state-of-the-art method that serves as a primary baseline for our work. This sophisticated network extracts high-level semantic features from EEG for visual decoding, representing the key performance benchmark we aim to surpass.

## D More Visualizations

**More Successful Examples of Reconstruction** This section provides additional visualization results to further demonstrate the robustness and generality of our proposed model. As shown in Figures 4 and 5, each example directly compares our model’s reconstruction (**Ours**) with the ground truth video (**GT**). These supplementary examples cover a diverse range of semantic categories, including:

- **Human-related Scenarios:** human scenes, musical activities, etc.
- **Nature-related Scenarios:** land animals, water animals, Plants, landscape, etc.
- **Man-made Objects:** Food, Transportation, Cities, etc.

Across these varied scenarios, our model consistently generates reconstructions that are highly faithful to the ground

truth videos in semantic content, color, and temporal coherence. This visual evidence, combined with the quantitative results in the main paper, validates our model’s robust capability for decoding complex dynamic visual experiences from non-invasive EEG signals.

**Some failure samples** Although our model performs well in most cases, analyzing its failure cases allows us to more clearly identify its limitations. As shown in Fig. 6, these limitations primarily fall into three categories:

- **Partial Frame Distortion:** In some cases, while the core object is correctly identified, parts of the frame or background still exhibit visual distortion. For example, in the “panda” scene, the reconstruction captured the basic shape of the panda, but as the perspective changes in the video, the subject undergoes morphological distortion in some frames.
- **Incoherent Object Motion:** In the “motorcycle” example, although the model generates semantically correct single-frame images, there are issues with inconsistency or mismatch with the real world in terms of the object’s direction of motion and movement pattern.
- **Incorrect Semantics:** In the “skiing” example, the model misinterprets a person on a blurry snow slope as someone skateboarding on a ramp. This suggests that for visually ambiguous inputs, the model can still converge to an incorrect semantic concept.

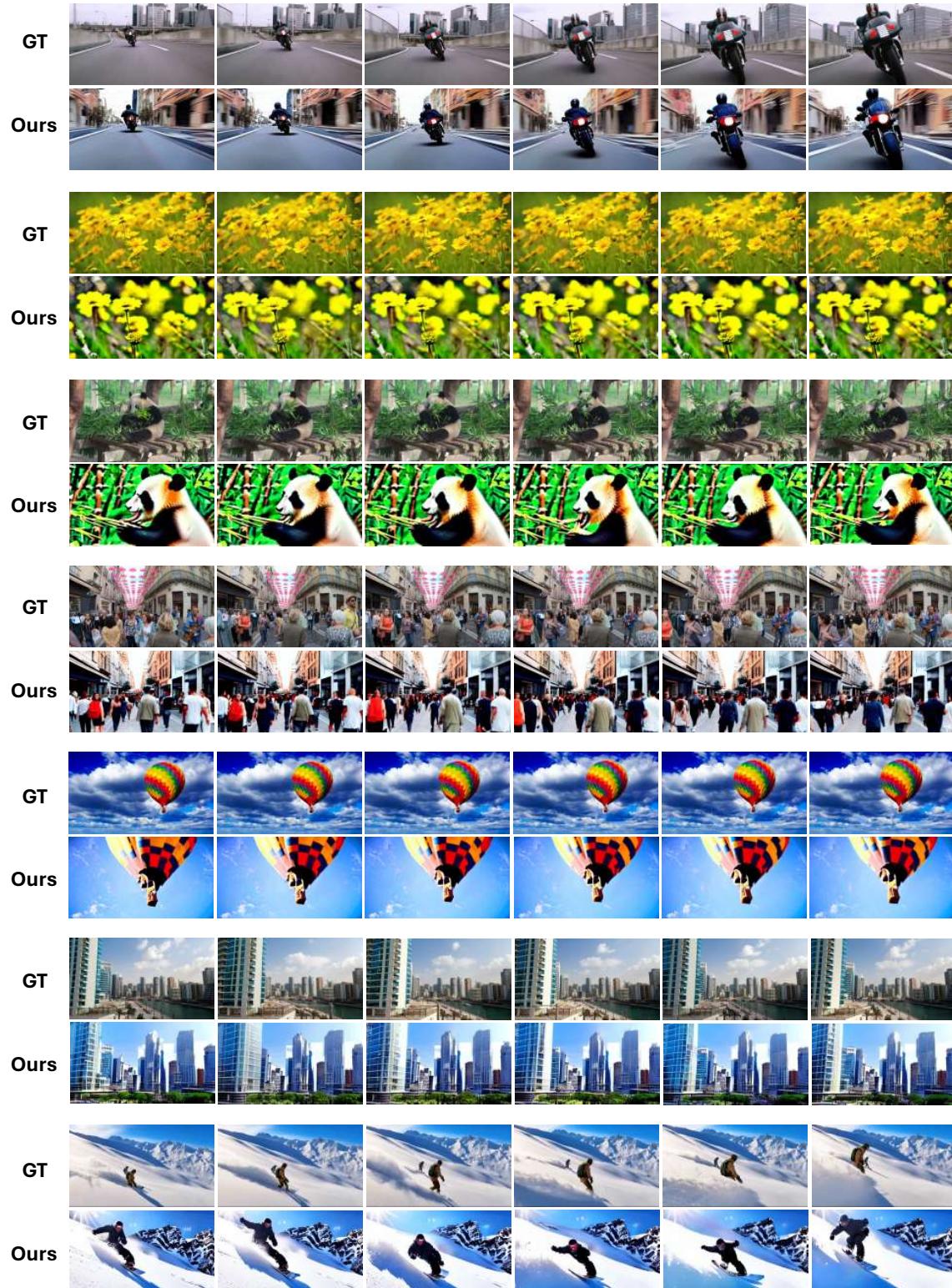


Figure 4: Various videos reconstruction samples.

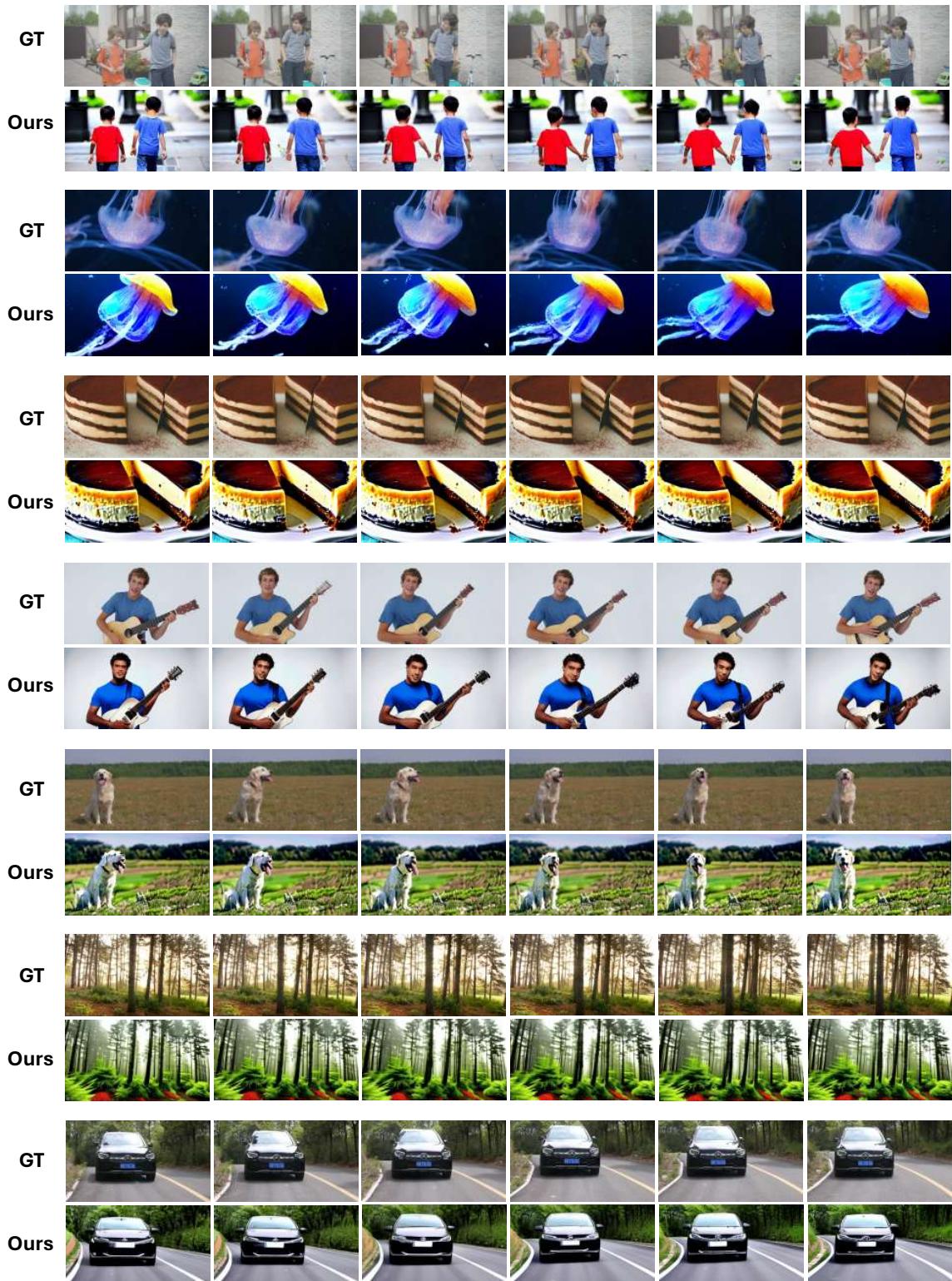


Figure 5: Various videos reconstruction samples.

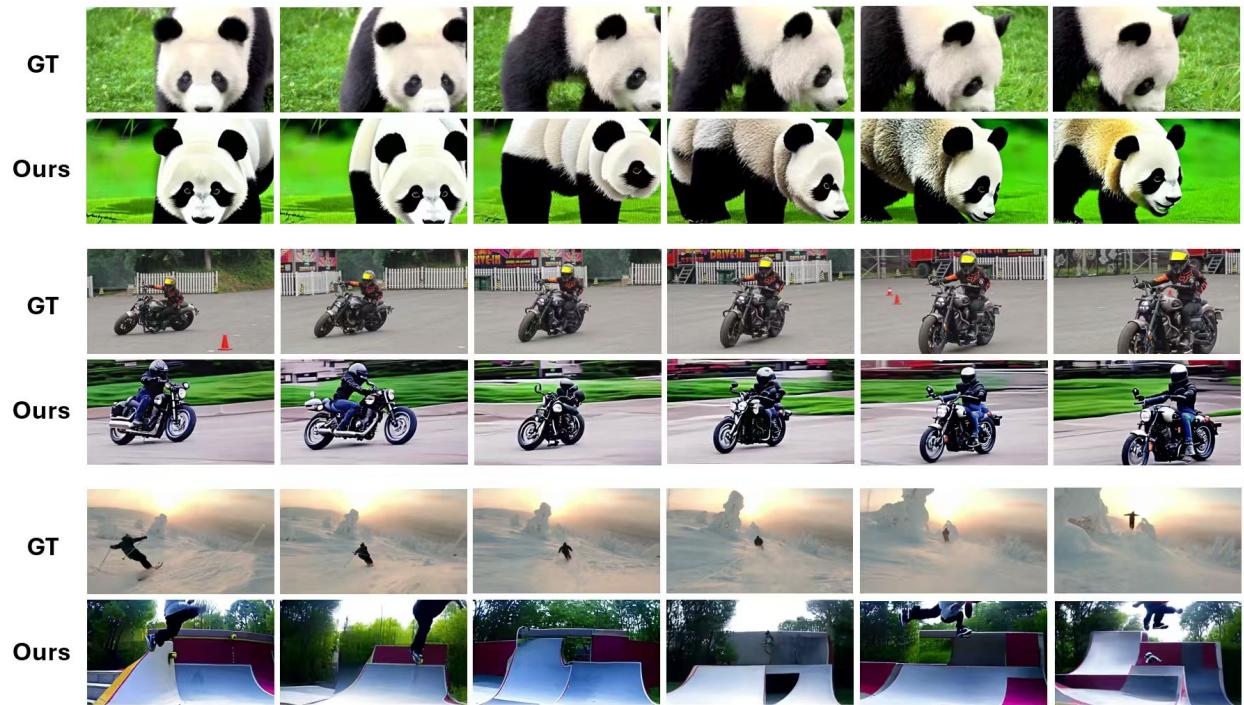


Figure 6: Some failure samples.